DENGUE FEVER IN SAN JUAN AND IQUITOS

ADEC7460 PREDICTIVE

ANALYTICS/FORECASTING

NICK HOWARD

08/26/2019

1 INTRODUCTION

The Dengue Virus

Each year up to 400 million people are infected with the dengue virus[1]. The dengue virus is a flavivirus that can cause severe infections such as Dengue Fever and Dengue Hemorrhagic Fever.

The Development of Dengue Hemorrhagic Fever

There are four subtypes of the dengue virus. The danger lies in multiple infections, if a person is infected with one type of dengue they typically have an immune response that clears up that infection. If they are infected with another type of dengue after the initial infection the immune system will react with the same immune response. This is because each virus subtype has similar surface antigens. The immune response is ineffective for the second subtype and the immune response will actually damage endothelial tissue which can result in hemmoraghic fever and fluid loss. The risk of death in those who develop Dengue Hemorrhagic fever is much higher than those who only have the dengue virus[3]. Once infected with the Dengue Virus a mother can spread this virus to a child in utero and through the act of breast feeding[2].

Benefit of stopping Initial and Secondary Infections

Since multiple infections can be very dangerous there is a benefit in stopping the initial infection through the deployment of vaccines to high risk areas. Stopping an initial outbreak will greatly reduce the risk of developing Dengue Fever/Dengue Hemorrhagic Fever. If there has been an outbreak of initial infections there is an even greater benefit of stopping subsequent infections in the infected population, doing so may save the lives of high-risk populations such as children, the elderly, and those with weak endothelial tissues.

2 DATA

2.1 FEATURE DESCRIPTIONS

The features provided to us described characteristics of the city's location, climate, and agricultural health. Along with the features described below, each observation includes the **year and weekofyear** variables to show when the observation occurred. The variable **total_cases** is also included to show the total number of diagnosed cases of Dengue Virus in each city on a given week. When applicable units are shows with the suffix **_units** in the features title (I.e **_temp** and **_mm**)

LOCATION

1. **City and date indicators**

city – City abbreviations: sj for San Juan and iq for Iquitos

week_start_date – Date given in yyyy-mm-dd format

CLIMATE

1. **NOAA's GHCN daily climate data weather station measurements**

station_max_temp_c – Maximum temperature

station_min_temp_c – Minimum temperature

station_avg_temp_c – Average temperature

station_precip_mm – Total precipitation

station_diur_temp_rng_c – Diurnal temperature range

2. **PERSIANN satellite precipitation measurements (0.25x0.25 degree scale)**

precipitation_amt_mm – Total precipitation

3. **NOAA's NCEP Climate Forecast System Reanalysis measurements (0.5x0.5 degree scale)**

reanalysis_sat_precip_amt_mm – Total precipitation

reanalysis_dew_point_temp_k – Mean dew point temperature reanalysis_air_temp_k – Mean air temperature

reanalysis_relative_humidity_percent – Mean relative humidity

reanalysis_specific_humidity_g_per_kg – Mean specific humidity

reanalysis_precip_amt_kg_per_m2 – Total precipitation

reanalysis_max_air_temp_k – Maximum air temperature

reanalysis_min_air_temp_k – Minimum air temperature

reanalysis_avg_temp_k – Average air temperature

reanalysis_tdtr_k – Diurnal temperature range

AGRICULTURE

1. Satellite vegetation - Normalized difference vegetation index (NDVI) –

NOAA's CDR Normalized Difference Vegetation Index (0.5x0.5 degree scale) measurements

ndvi_se – Pixel southeast of city centroid

ndvi_sw – Pixel southwest of city centroid

ndvi_ne – Pixel northeast of city centroid

ndvi_nw – Pixel northwest of city centroid

2.2 DESCRIPTIVE STATISTICS

str() Function

24 Variables. Most of the variables are measurements (continuous) and there are a few factors. The total cases were in another dataset and merged into this dataset after the initial data exploration.

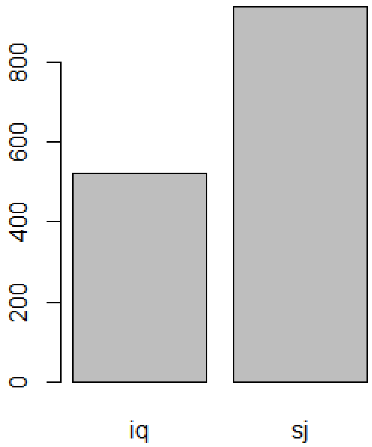
```
 data.frame':  1456 obs. of  24 variables:
  $ city                                : Factor w/ 2 levels "iq","sj": 2 2 2 2 2 2 2 2 2 2 ...
  $ year                                : int  1990 1990 1990 1990 1990 1990 1990 1990 1990 1990 ...
  $ weekofyear                          : int  18 19 20 21 22 23 24 25 26 27 ...
  $ week_start_date                     : Factor w/ 1049 levels "1990-04-30","1990-05-07",..: 1 2 3 4 5 6 7 8 9 10 ...
  $ ndvi_ne                             : num  0.1226 0.1699 0.0323 0.1286 0.1962 ...
  $ ndvi_nw                             : num  0.104 0.142 0.173 0.245 0.262 ...
  $ ndvi_se                             : num  0.198 0.162 0.157 0.228 0.251 ...
  $ ndvi_sw                             : num  0.178 0.155 0.171 0.236 0.247 ...
  $ precipitation_amt_mm                : num  12.42 22.82 34.54 15.36 7.52 ...
  $ reanalysis_air_temp_k               : num  298 298 299 299 300 ...
  $ reanalysis_avg_temp_k               : num  298 298 299 299 300 ...
  $ reanalysis_dew_point_temp_k         : num  292 294 295 295 296 ...
  $ reanalysis_max_air_temp_k           : num  300 301 300 301 302 ...
  $ reanalysis_min_air_temp_k           : num  296 296 297 297 298 ...
  $ reanalysis_precip_amt_kg_per_m2     : num  32 17.9 26.1 13.9 12.2 ...
  $ reanalysis_relative_humidity_percent : num  73.4 77.4 82.1 80.3 80.5 ...
  $ reanalysis_sat_precip_amt_mm        : num  12.42 22.82 34.54 15.36 7.52 ...
  $ reanalysis_specific_humidity_g_per_kg: num  14 15.4 16.8 16.7 17.2 ...
  $ reanalysis_tdtr_k                   : num  2.63 2.37 2.3 2.43 3.01 ...
  $ station_avg_temp_c                  : num  25.4 26.7 26.7 27.5 28.9 ...
  $ station_diur_temp_rng_c             : num  6.9 6.37 6.49 6.77 9.37 ...
  $ station_max_temp_c                  : num  29.4 31.7 32.2 33.3 35 34.4 32.2 33.9 33.9 33.9 ...
  $ station_min_temp_c                  : num  20 22.2 22.8 23.3 23.9 23.9 23.3 22.8 22.8 24.4 ...
  $ station_precip_mm                   : num  16 8.6 41.4 4 5.8 39.1 29.7 21.1 21.1 1.1 ...
```

SUMMARY STATISTICS: FULL DATASET

CITY

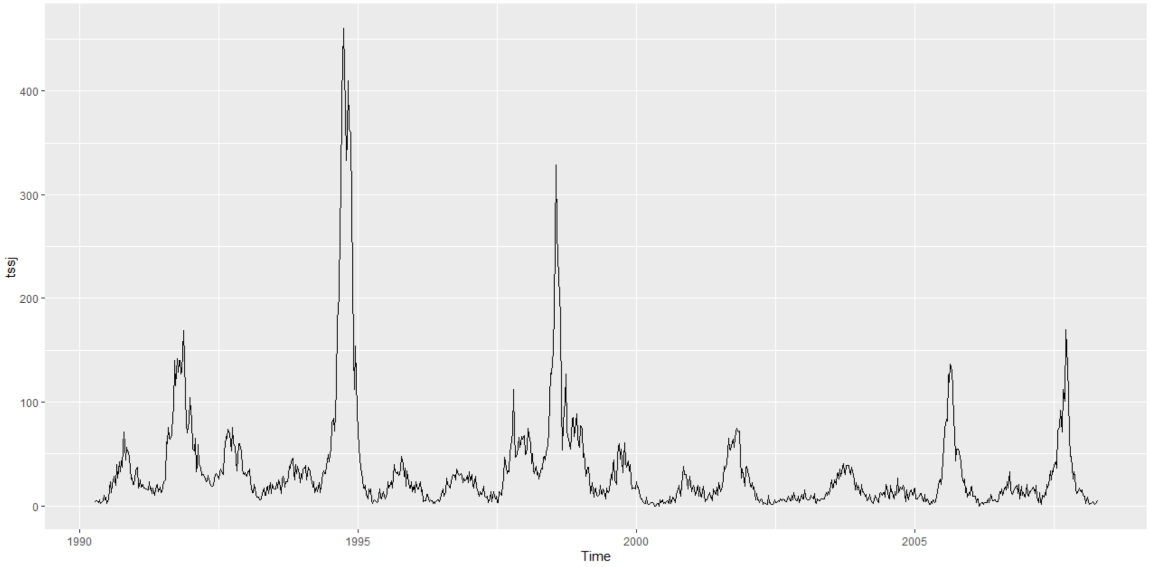
There was a total of 936 weekly recordings of the rate of dengue fever in San Juan and there were 520 recordings in Iquitos.

San Juan	Iquitos
936	520



SUMMARY STATISTICS: SAN JUAN

To get a better idea of the impact of the Dengue Fever in each city I created two subsets of death. One subset contains observations from the city of San Juan (**Denguesj**) and the second subset contains observations from the city of Iquitos (**Dengueiq**).



TIME

The observations were recorded weekly from 1990 to 2008. Observations started on the week of 04/30 in 1990.

```
year
Min.   :1990
1st Qu.:1994
Median :1999
Mean   :1999
3rd Qu.:2003
Max.   :2008
```

TOTAL CASES OF DENGUE VIRUS

Over the 936 weeks that rates of the Dengure Virus were recorded the highest number of observations was 461. The median number of observations was 19 and the mean was 37 cases.

```
 Min.   :  0.00
 1st Qu.:  9.00
 Median : 19.00
 Mean   : 34.18
 3rd Qu.: 37.00
 Max.   :461.00
```

CLIMATE - Denguesj

The mean average max temperature in Celsius was 31.61 degrees and the mean minimum temperature in Celsius was 22.6 degress.

```
station_max_temp_c station_min_temp_c
 Min.   :26.70      Min.   :17.8      
 1st Qu.:30.60      1st Qu.:21.7      
 Median :31.70      Median :22.8      
 Mean   :31.61      Mean   :22.6      
 3rd Qu.:32.80      3rd Qu.:23.9      
 Max.   :35.60      Max.   :25.6      
 NA's   :6          NA's   :6         
```

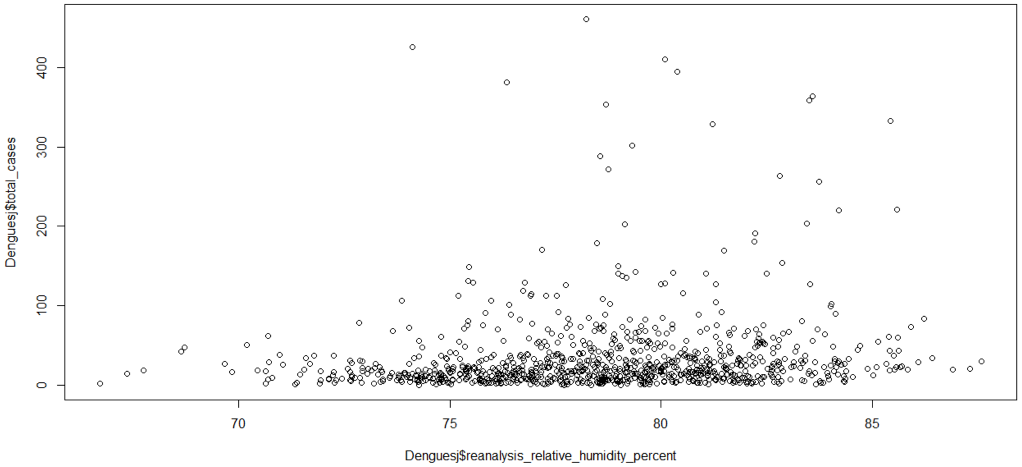
The total precipation in San Juan from 1990 to 2008 was 26 millimeters. In the rainy season this region can get up to 306 milimeters of rainfall per week

.

```
station_precip_mm
 Min.   :  0.000  
 1st Qu.:  6.825  
 Median : 17.750  
 Mean   : 26.785  
 3rd Qu.: 35.450  
 Max.   :305.900  
```

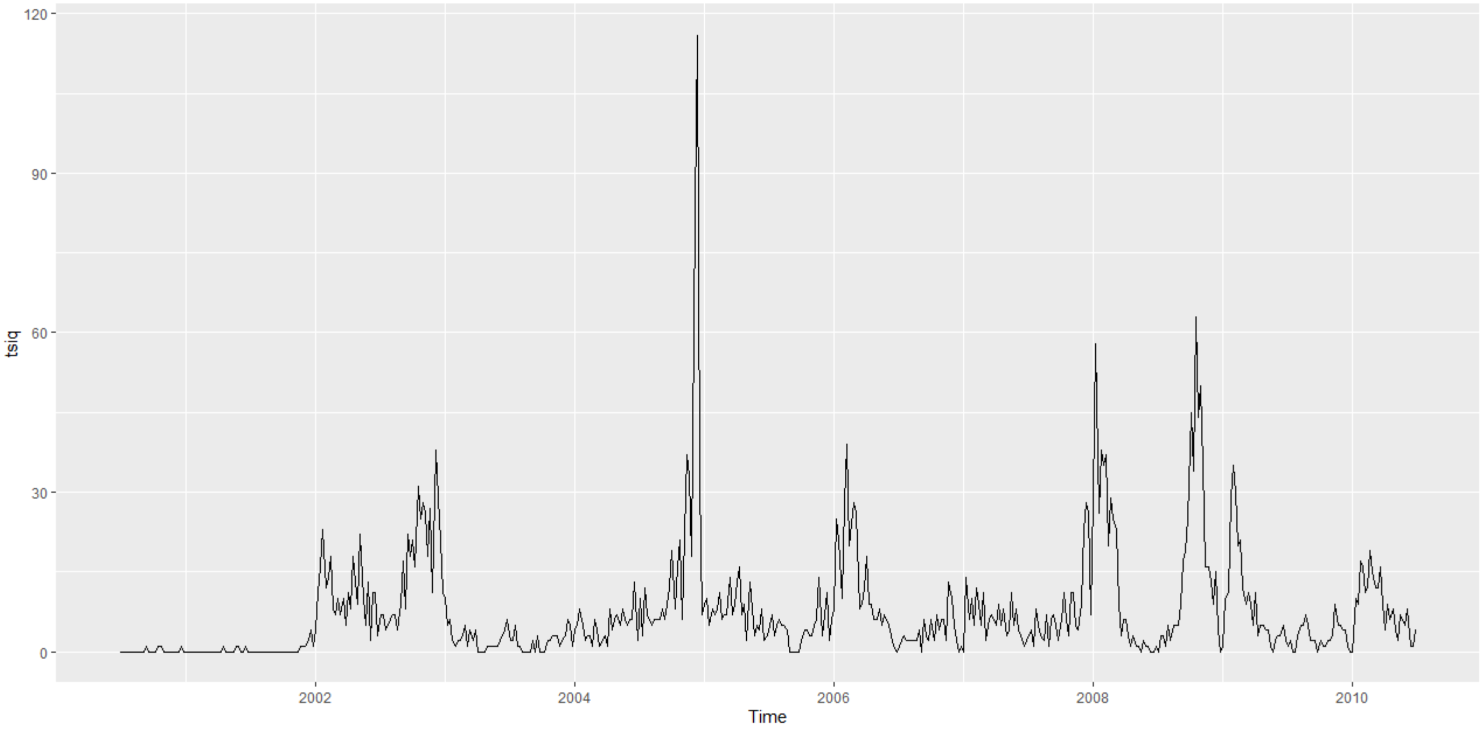
This region is humid. The relative humidity percent could be up to 87%. As humidity increased past 75% the observed cases of Dengue Fever increased.

```
reanalysis_relative_humidity_percent
 Min.   :66.74
 1st Qu.:76.25
 Median :78.67
 Mean   :78.57
 3rd Qu.:80.96
 Max.   :87.58
 NA's   :6
```



SUMMARY STATISTICS: IQUITOS

To get a better idea of the impact of the Dengue Fever in each city I created two subsets of death. One subset contains observations from the city of San Juan (**Denguesj**) and the second subset contains observations from the city of Iquitos (**Dengueiq**).



TIME

The observations were recorded weekly from 2000 to 2010. Observations started on the week of 04/30 in 1990.

```
    year     
Min.   :2000  
1st Qu.:2003  
Median :2005  
Mean   :2005  
3rd Qu.:2007  
Max.   :2010  
```

TOTAL CASES OF DENGUE VIRUS

Over the 520 weeks that rates of the Dengue Virus were recorded the highest number of observations was 116. The median number of observations was 5 and the mean was 8 cases.

```
  total_cases    
 Min.   :  0.000  
 1st Qu.:  1.000  
 Median :  5.000  
 Mean   :  7.565  
 3rd Qu.:  9.000  
 Max.   :116.000  
```

CLIMATE - Iquitos

The mean average max temperature in Celsius was 34 degrees and the mean minimum temperature in Celsius was 21 degrees Celsius.

```
station_max_temp_c station_min_temp_c
 Min.   :30.1       Min.   :14.7      
 1st Qu.:33.2       1st Qu.:20.6      
 Median :34.0       Median :21.3      
 Mean   :34.0       Mean   :21.2      
 3rd Qu.:34.9       3rd Qu.:22.0      
 Max.   :42.2       Max.   :24.2      
 NA's   :14         NA's   :8         
```

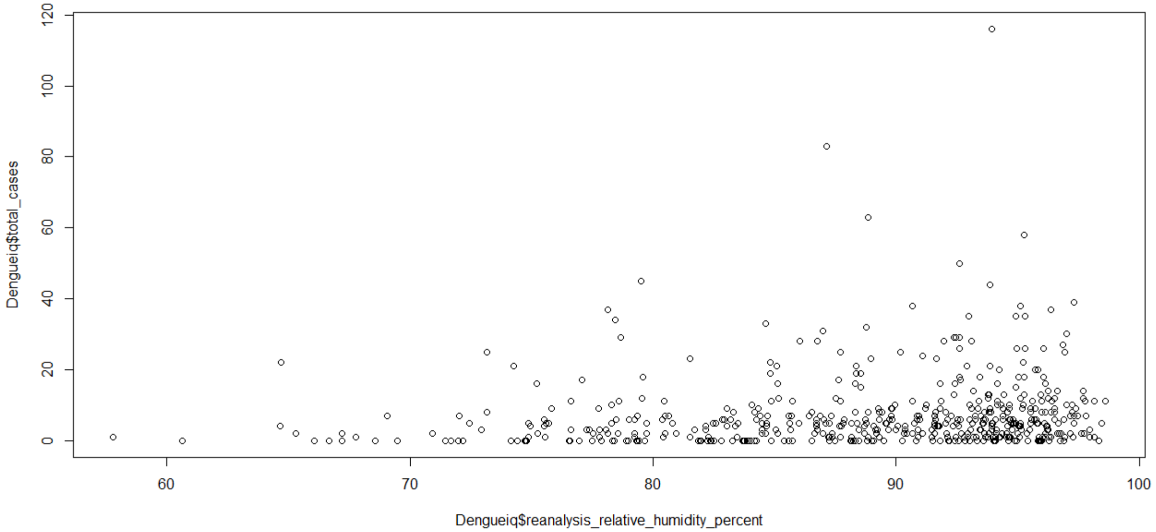
The total precipation in San Juan from 1990 to 2008 was 26 millimeters. In the rainy season this region can get up to 306 milimeters of rainfall per week

.

```
station_precip_mm
 Min.   :  0.00
 1st Qu.: 17.20
 Median : 45.30
 Mean   : 62.47
 3rd Qu.: 85.95
 Max.   :543.30
 NA's   :16
```

This region is very humid. The relative humidity percent could be up to 99%. As humidity increased past 75% the observed cases of Dengue Fever increased.

```
reanalysis_relative_humidity_percent
 Min.   :57.79
 1st Qu.:84.30
 Median :90.92
 Mean   :88.64
 3rd Qu.:94.56
 Max.   :98.61
 NA's   :4
```

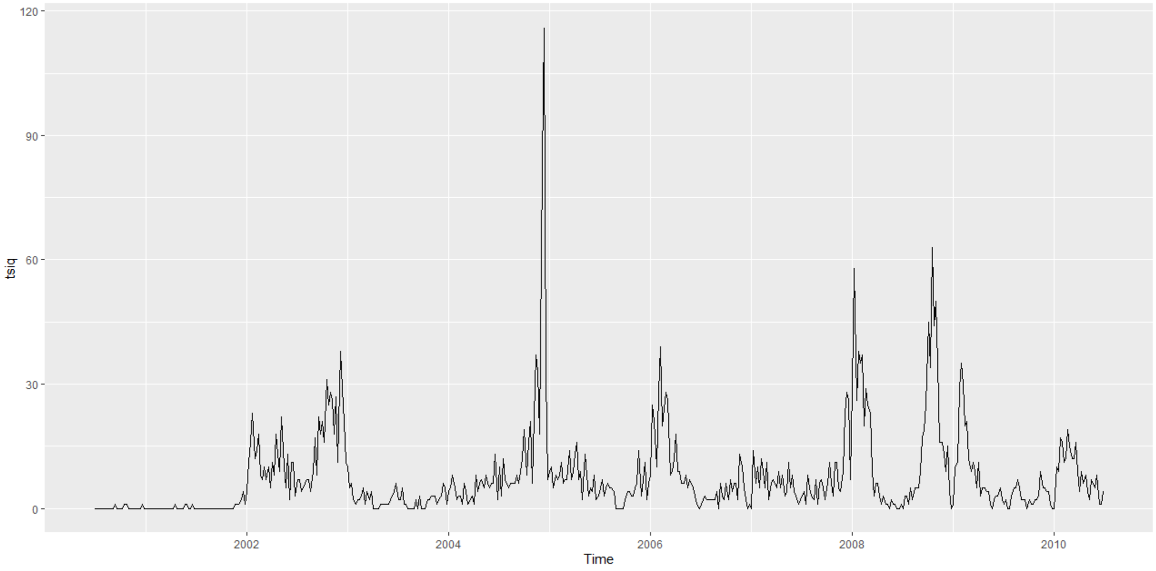


3 FORECASTS: IQUITOS

For this assignment I will be using the Forecast package to complete forecasts on the Iquitos subset of the full Dengue Virus dataset.

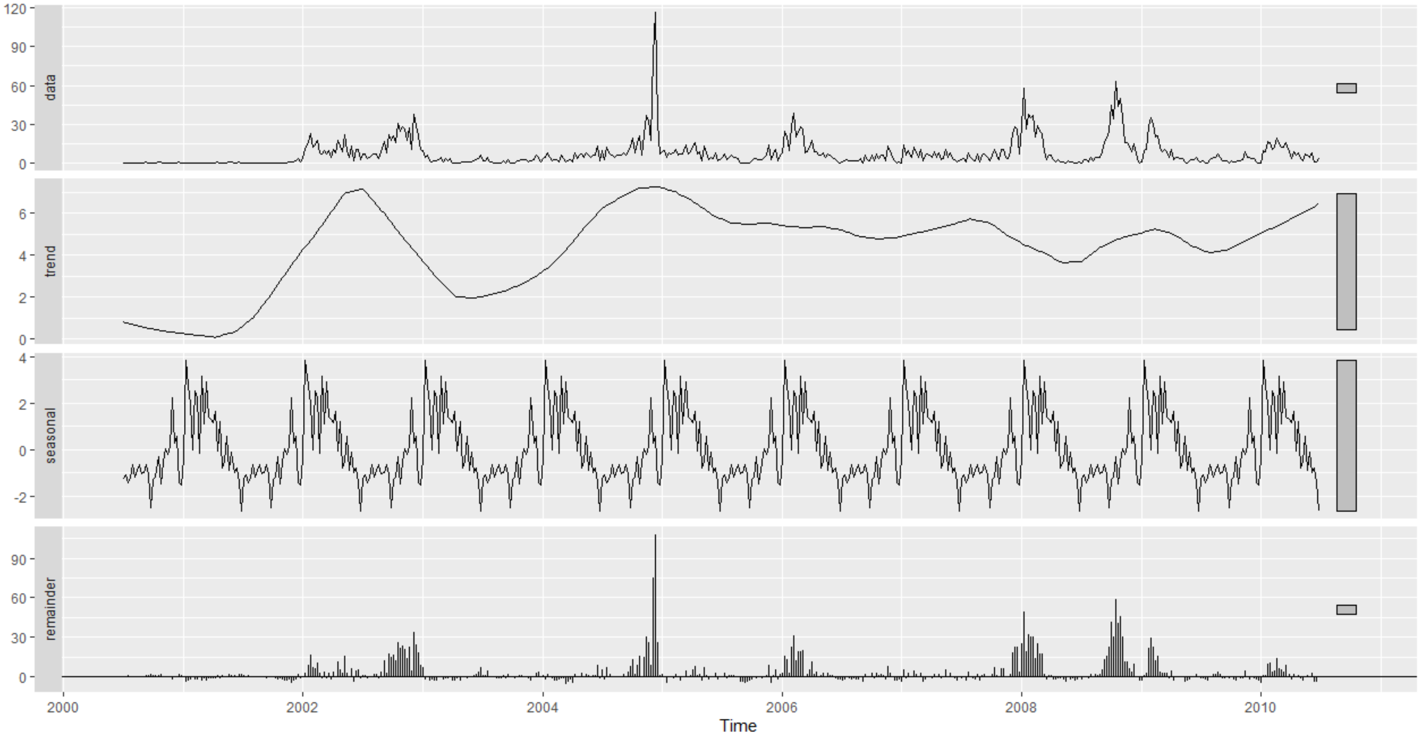
TIME SERIES

- The first step I took was to create a time series object of the **Dengueiq** subset.
- The frequency parameters was set to 52 (weekly). The start was in week 27 of 2000.



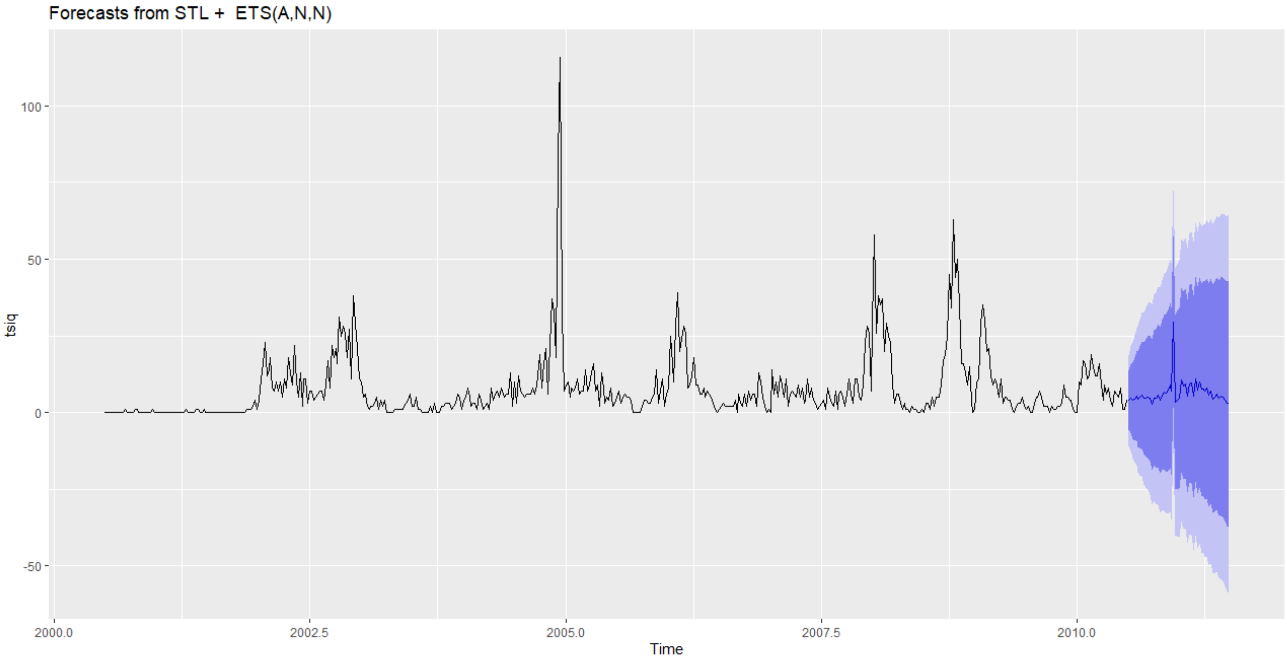
DECOMPOSITION – STL

To decompose this data I used the Seasonal and Trend decomposition using Loess method (STL). This method will handle seasonality of weekly ts data. There was a strong seasonal component to this data and the trend is mostly positive until it levels out in 2006.



4.1 STLF()

The first model that I used to forecast was the **stlf()** method. This method has the advantage of being able to handle time series data with high frequencies (weekly). The function arguments were set to their default values.

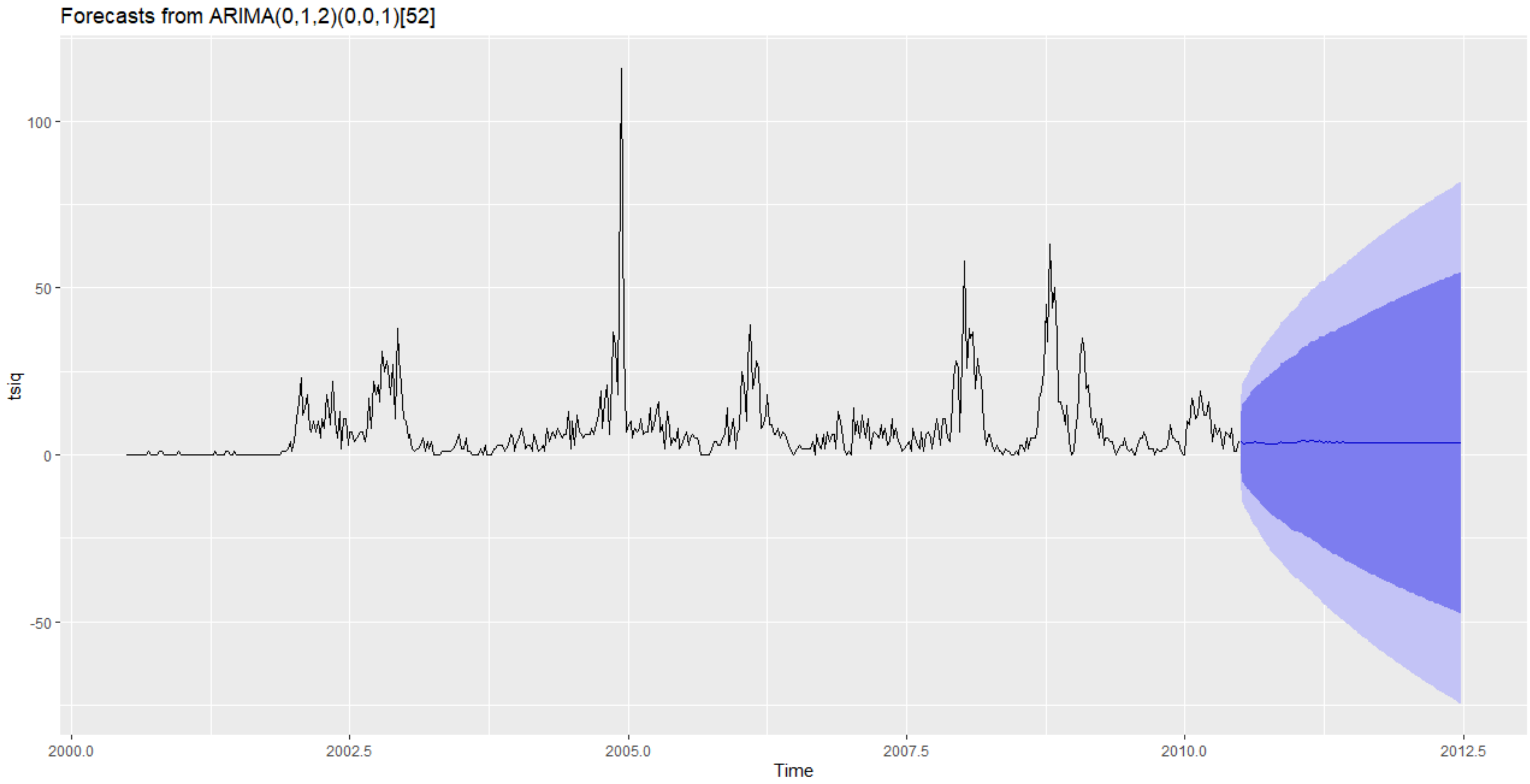


ACCURACY STLF

The RMSE will be used to judge the model’s accuracy. The RMSE of this model was 0.9751016.

AUTO.ARIMA()

The second model that I chose to use was the ARIMA Model. ARIMA Provides a complementary approach to an exponential smoothing method. While the Exp. Smoothing models are based on the trend and seasonality that is present in the data, the ARIMA family of models attempt to describe autocorrelations.

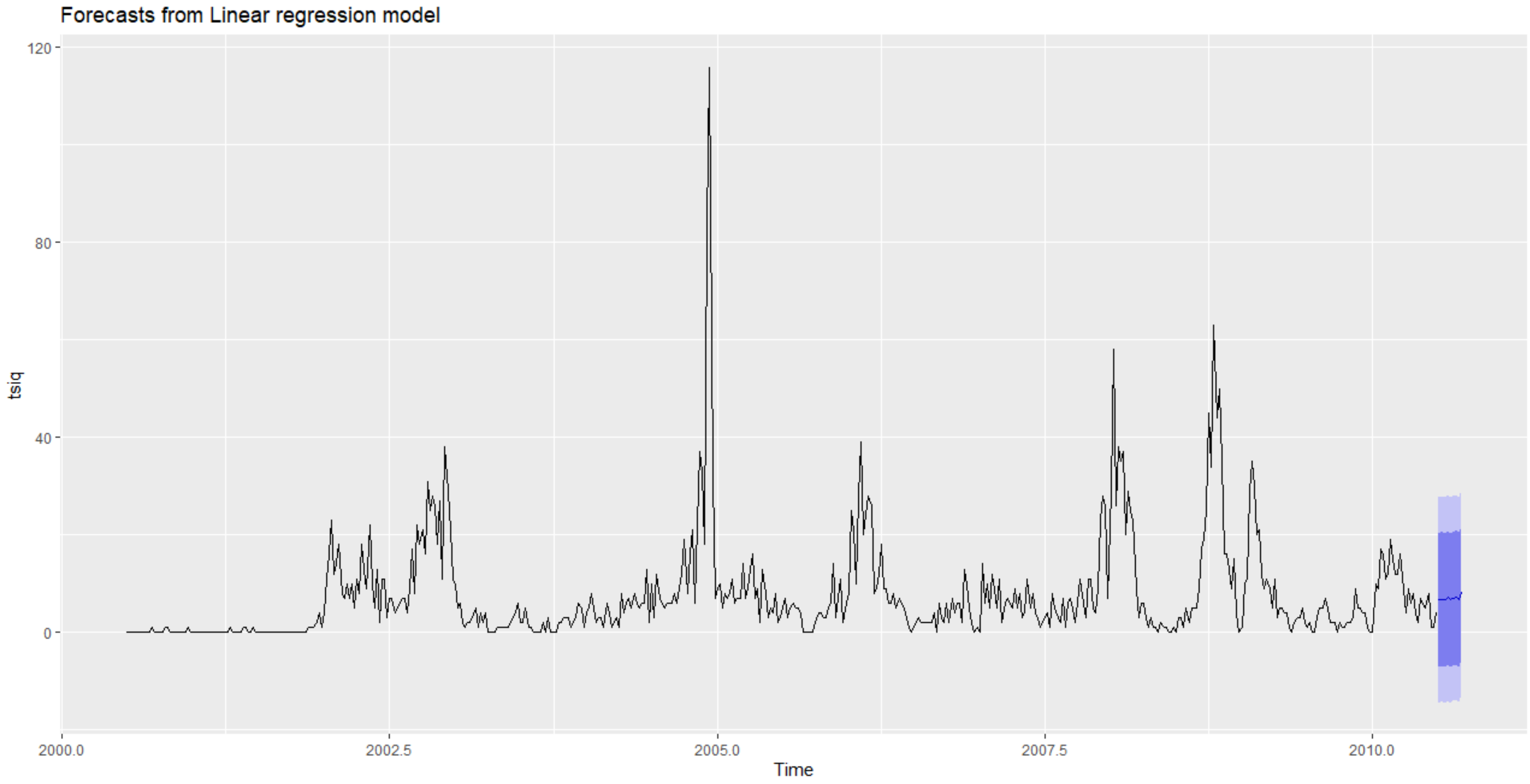


ACCURACY AUTO.ARIMA

The RMSE will be used to judge the model's accuracy. The RMSE of this model was 7.191928**.**

TSLM

The last model I used was linear regression for time series data.



ACCURACY TSLM

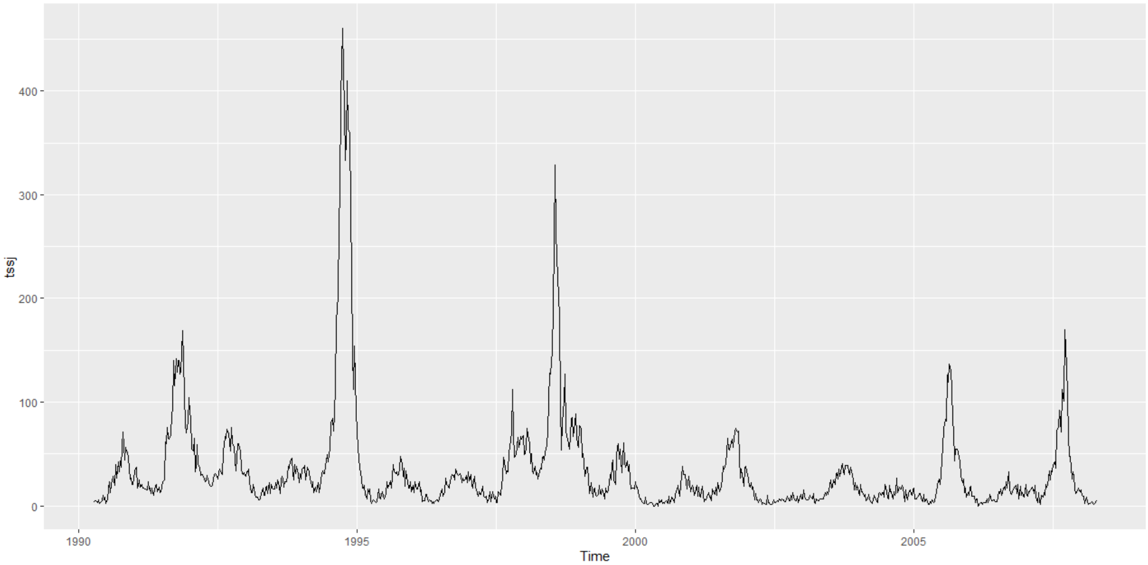
The RMSE will be used to judge the model’s accuracy. The RMSE of this model was 9.634941**.**

4 FORECASTS: SAN JUAN

For this assignment I will be using the Forecast package to complete forecasts on the San Juan subset of the full Dengue Virus dataset.

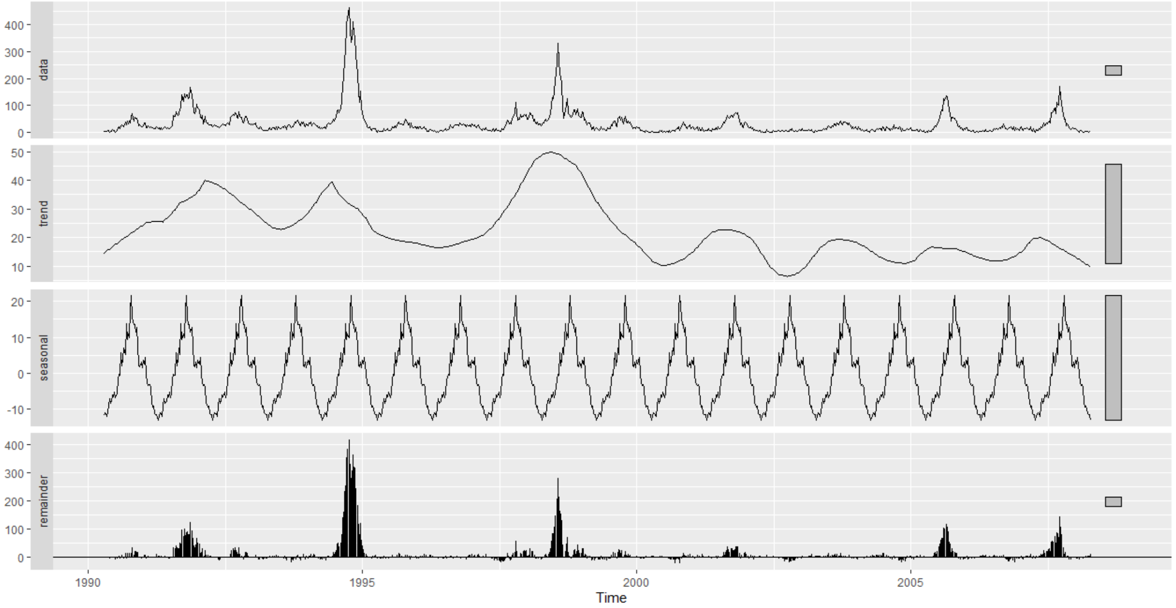
TIME SERIES

- The first step I took was to create a time series object of the **Denguesj** subset.
- The frequency parameters was set to 52 (weekly). The start was in week 16 of 1990.



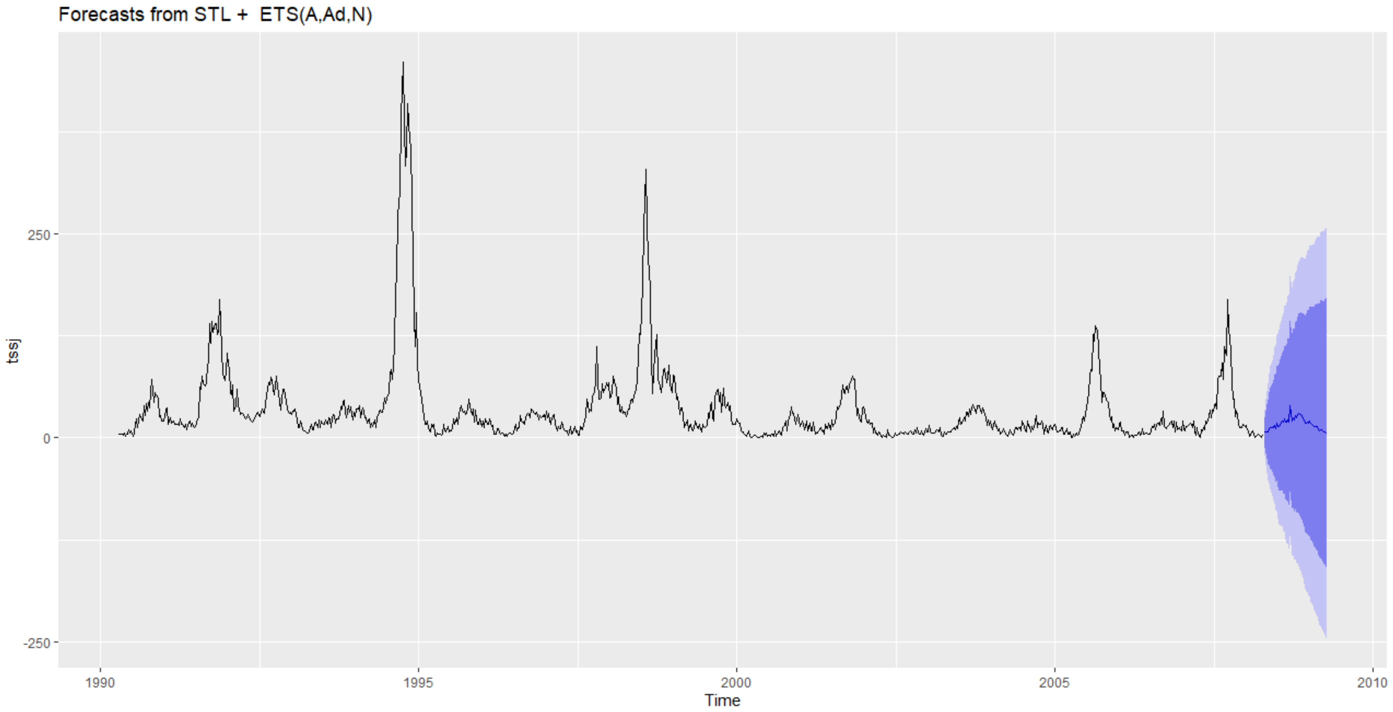
DECOMPOSITION – STL

To decompose this data I used the Seasonal and Trend decomposition using Loess method (STL). This method will handle seasonality of weekly ts data. There was a strong seasonal component to this data and there consistent variation between positive upward trends and downward trends in the data.



STLF()

The first model that I used to forecast was the **stlf()** method. This method has the advantage of being able to handle time series data with high frequencies (weekly). The arguments were set to their default values.

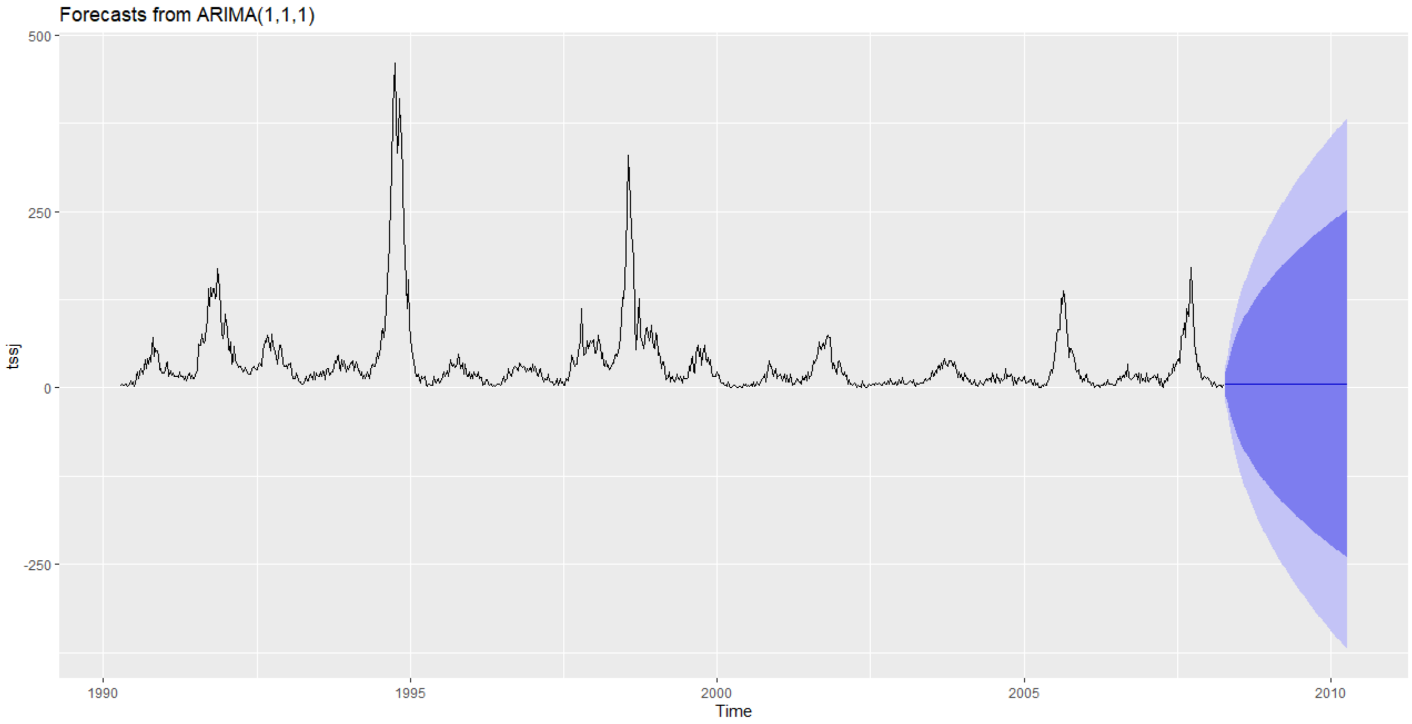


ACCURACY STLF

The RMSE will be used to judge the model’s accuracy. The RMSE of this model was **0.6399651**.

AUTO.ARIMA()

The second model that I chose to use was the ARIMA Model. ARIMA Provides a complementary approach to an exponential smoothing method. While the Exp. Smoothing models are based on the trend and seasonality that is present in the data, the ARIMA family of models attempt to describe autocorrelations. The arguments were set to their default values.

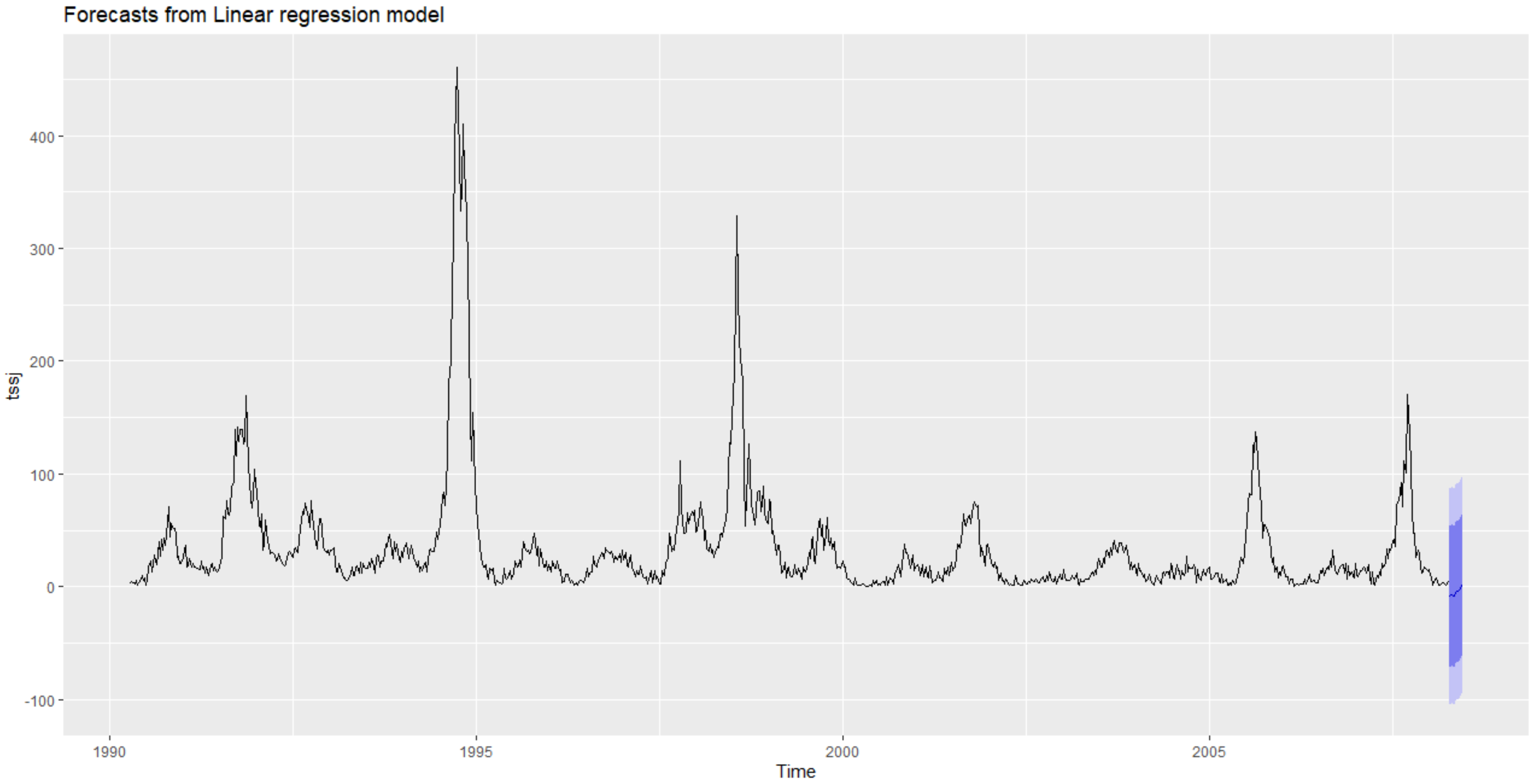


ACCURACY AUTO.ARIMA

The RMSE will be used to judge the model’s accuracy. The RMSE of this model was **13.42959.** This is very high and is not a good sign for predictive accuracy.

TSLM

The last model I used was linear regression for time series data.



ACCURACY TSLM

The RMSE will be used to judge the model’s accuracy. The RMSE of this model was **46.02154.**This is very high and is not a good sign for predictive accuracy.

5 PERFORMANCE

For both subsets of data (San Juan and Iquitos) the STLF method provided the best model on the metric of RMSE. The ARIMA model provided the second best model in both cases. The ARIMA model did not perform nearly as well as the STLF Model. The worst performing model was the linear model.

Model	Iquitos RMSE	San Juan RMSE
STLF	**0.9751016**	**0.6399651**
ARIMA	7.191928	13.42959
LM	9.634941	46.02154

DISCUSSION OF PERFORMANCE AND LIMITATIONS

STLF

I chose to use the STLF model after initially attempting to use the exponential smoothing function **ets().** R advised me to use STLF() due to the high frequency (52 for weekly data). This model assumes that the timeseries can be broken down into error, trend, and seasonality components[5]. We saw in the STL forecast that this was true. Since this data had a predictable seasonal component STLF was a very good choice and this is reflected in the low RMSE.

ARIMA

The ARIMA model produced a decent forecast based on the metrics of RMSE. The method I chose to use was the Auto.Arima() method. This method chooses the best model based on minimizing the AIC. This model is appropriate for stationary time series data.

This means that this model is intended to be used when the mean, variance, and autocorrelations are constant[5]. This data could not be described as stationary and an ARIMA model was probably not the best choice.

LINEAR MODEL

The Linear Model can be used for modeling non-linear data. For this forecasts I used the tslm() function in R. I used the default values for the function arguments. Lambda (for transformations) was not specified so that transformation was left up to R. I used trend and season to model this data. In the future I would like to experiment with piecewise linear regression models by adding knots and regression splines.

LEARNING AND FUTURE WORK

I learned quite a bit about the impact of Dengue Fever and the consequences that this virus can have for those infected. For my modeling work I learned that the ETS model is not appropriate for timeseries data with a high frequency, that the ARIMA model works best with stationary data, and that the linear model is not a great predictive model to timeseries data unless it is augmented through the use of piecewise regression tools such as knots or smoothing splines. In the future I would like to spend more time on the STL model and experiment with the function arguments to refine the model. I would like to experiment with variable transformations as well. If there is a way to transform the data to create a stationary time series I think this would be useful so I can then perform ARIMA. Finally I would like to review the piecewise regression methods to see if I can improve my linear models for time series data.

Appendix 1 - Works Cited

1) Centers for Disease Control and Prevention (n.d). Dengue. Retrieved from https://www.cdc.gov/dengue/prevention/index.html
2) Bernke te Winkel and Christof Schaefer (2015). 4.15 - Infections during breastfeeding. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780124080782000408
3) Stanford (n.d). **Dengue Virus Profile**. Retrieved from http://web.stanford.edu/group/virus/flavi/2000/dengue.htm
4) Mohit Mayank (2018, May 20). Understanding the forecasting algorithm: STLF Model. Retrieved from https://itnext.io/understanding-the-forecasting-algorithm-stlf-model-29d74b3a0336
5) STATISTICA (n.d). **ARIMA - Evaluation of the Model**. Retrieved from https://documentation.statsoft.com/STATISTICAHelp.aspx?path=TimeSeries/TimeSeries/Overview/Arima/ARIMAEvaluationoftheModel

Appendix 2 – R Code

```
library(ggplot2)
library(fpp2)
library(forecast)
library(ggplot2)
library(tseries)


Dengue<- read.csv("C:/Users/Nicholas Howard/Desktop/Applied Economics/Forecasting/final paper/DengAI_Predicting_Disease_Spread_-_Training_Data_Features.csv")


str(Dengue)


summary(Dengue)
summary(Dengueiq)
summary(Denguesj)


Denguesj<-subset(Dengue, Dengue$city=="sj")
Dengueiq<-subset(Dengue, Dengue$city=="iq")
```

```
plot(Dengueiq$reanalysis_relative_humidity_percent,Dengueiq$total_cases)

plot(Denguesj$reanalysis_relative_humidity_percent,Denguesj$total_cases)

tssj<-ts(Denguesj$total_cases, frequency=52, start = c(1990,16)

autoplot(decompose(Denguesj, type="additive"))

autoplot(decompose(Denguesj, type="multiplicative"))

stl1<-tssj %>%
stl(t.window-13, s.window="periodic", robust=TRUE)%>%
autoplot()

stlf1<-stlf(tssj, h = 52, s.window = 13, t.window = NULL,
  robust = TRUE, lambda = NULL, biasadj = FALSE, PI=FALSE)

autoplot(forecast(stlf(tssj, h = 52, s.window = 13, t.window = NULL,
  robust = TRUE, lambda = NULL, biasadj = FALSE, PI=FALSE))
```

```
accuracy(stlf1)

arima1<-auto.arima(tssj)

autoplot(forecast(arima1),PI=FALSE)

accuracy(arima1)

lm1<-tslm(tssj~trend+season)

autoplot(forecast(lm1))


tsiq<-ts(Dengueiq$total_cases, frequency=52, start = c(1990,16)

autoplot(decompose(Dengueiq, type="additive"))

autoplot(decompose(Dengueiq, type="multiplicative"))
```

```
stl2<-tsiq %>%
stl(t.window-13, s.window="periodic", robust=TRUE)%>%
autoplot()

stlf2<-stlf(tsiq, h = 52, s.window = 13, t.window = NULL,
  robust = TRUE, lambda = NULL, biasadj = FALSE, PI=FALSE)

autoplot(forecast(stlf(tsiq, h = 52, s.window = 13, t.window = NULL,
  robust = TRUE, lambda = NULL, biasadj = FALSE,)))

accuracy(stlf2)

arima2<-auto.arima(tsiq)

autoplot(forecast(arima2),PI=FALSE)

accuracy(arima2)

lm2<-tslm(tsiq~trend+season)
```

```
autoplot(forecast(lm2))


accuracy(lm2)


accuracy(stlf1)
accuracy(stlf2)
accuracy(arima1)
accuracy(arima2)
accuracy(lm1)
accuracy(lm2)
```